Information Retrieval

Evaluation in IR systems

Yao-Chung Fan

yfan@nchu.edu.tw

National Chung Hsing University

Why System Evaluation?

- There are many retrieval models/ algorithms/ systems, which one is the best?
- What is the best component for:
  - Ranking function (dot-product, cosine, …)
  - Term selection (stopword removal, stemming…)
  - Term weighting (TF, TF-IDF,…)

Difficulties in Evaluating IR Systems

沒有標準答案、難量化

- Effectiveness is related to the ***relevancy*** of retrieved items.
- Relevancy is not typically binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- Relevancy, from a human standpoint, is:
  - Subjective: Depends upon a specific user's judgment.
  - Situational: Relates to user's current needs.
  - Cognitive: Depends on human perception and behavior.
  - Dynamic: Changes over time.

Human Labeled Corpora (Gold Standard)

- Start with a corpus of documents. 標記 data
- Collect a set of queries for this corpus.
- Have one or more human experts exhaustively label the relevant documents for each query.
- Typically assumes binary relevance judgments.
- Requires considerable human effort for large document/query corpora.

Standard Collections

TABLE 4.3 Common Test Corpora

Collection	*NDocs*	*NQrys*	*Size (MB)*	*Term/Doc*	*Q-D RelAss*
ADI	82	35			
AIT	2109	14	2	400	>10,000
CACM	3204	64	2	24.5	
CISI	1460	112	2	46.5	
Cranfield	1400	225	2	53.1	
LISA	5872	35	3		
Medline	1033	30	1		
NPL	11,429	93	3		
OSHMED	34,8566	106	400	250	16,140
Reuters	21,578	672	28	131	
TREC	740,000	200	2000	89-3543	» 100,000

每年一次

EVALUATING UNRANKED RESULTS

Precision and Recall

$\frac{a}{a+b}$

查準率

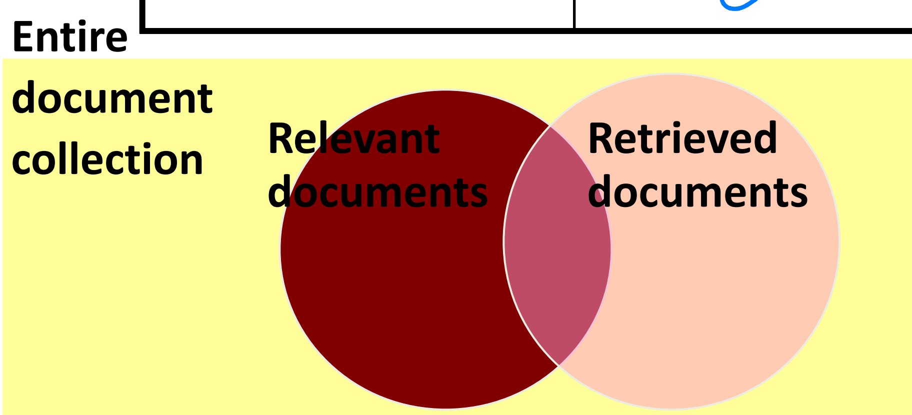
- **Precision**: fraction of retrieved docs that are relevant = P(relevant|retrieved) 抓了多少回來有多少相關

查全率

- **Recall**: fraction of relevant docs that are retrieved = P(retrieved|relevant) 針對查詢有多少相關 ⇒ 分母. 找到有多少 ⇒ 分母
  all num of related docs

$\frac{a}{a+c}$

	Relevant	Nonrelevant
Retrieved	tp a	fp b
Not Retrieved	fn c	tn d



n)

Recall 重要:
法條前案檢索
專利檢索

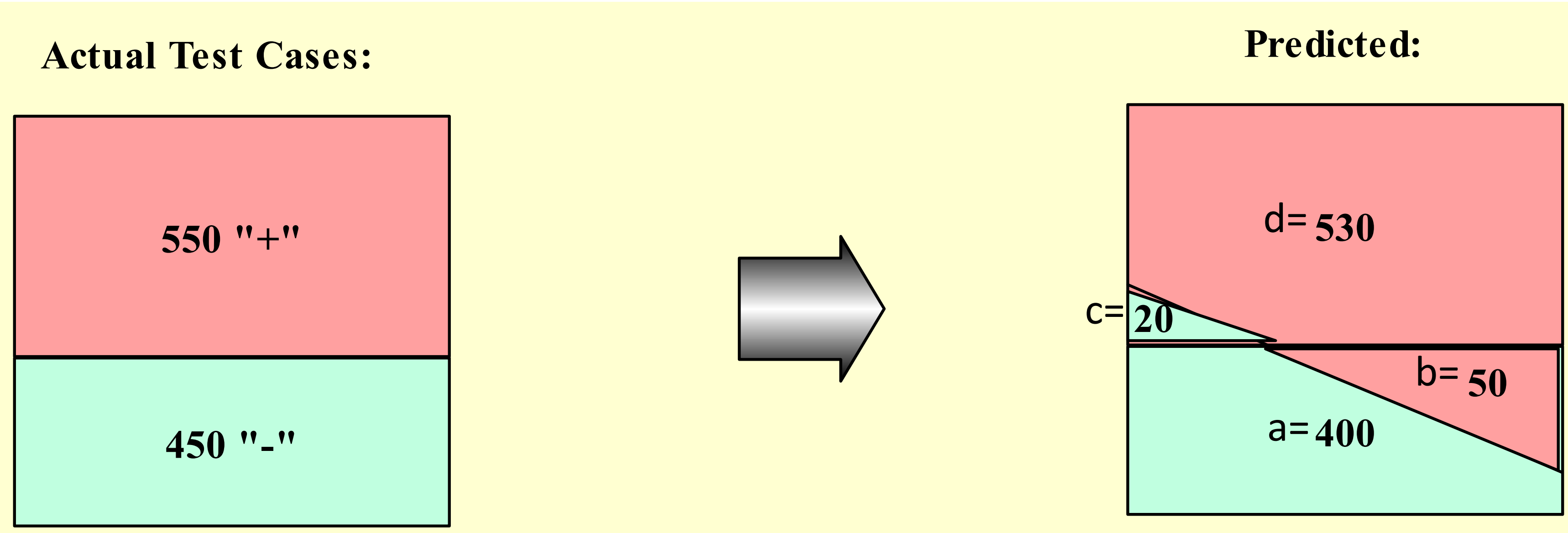
Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = P(relevant|retrieved)
- **Recall**: fraction of relevant docs that are retrieved = P(retrieved|relevant)

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)
- Accuracy A = (tp+tn)/(tp+fp+fn+tn)

Precision, Recall, Accuracy



For this :
a = 400
b = 50
c = 20
d = 530

Precision = d / (b + d) = 530 / 580 = 91.4%

Recall = d / (c + d) = 530 / 550 = 96.4%

Accuracy = (a+d)/(a+b+c+d)
= (530+400)/(530+20+50+400) = 93%

Should we instead use the accuracy measure for evaluation?

- Given a query, an engine classifies each doc as "Relevant" or "Nonrelevant"
- The **accuracy** of an engine: the fraction of these classifications that are correct
- **Accuracy** is a commonly used evaluation measure in machine learning classification work but not in IR

- Why is this not a very useful evaluation measure in IR?

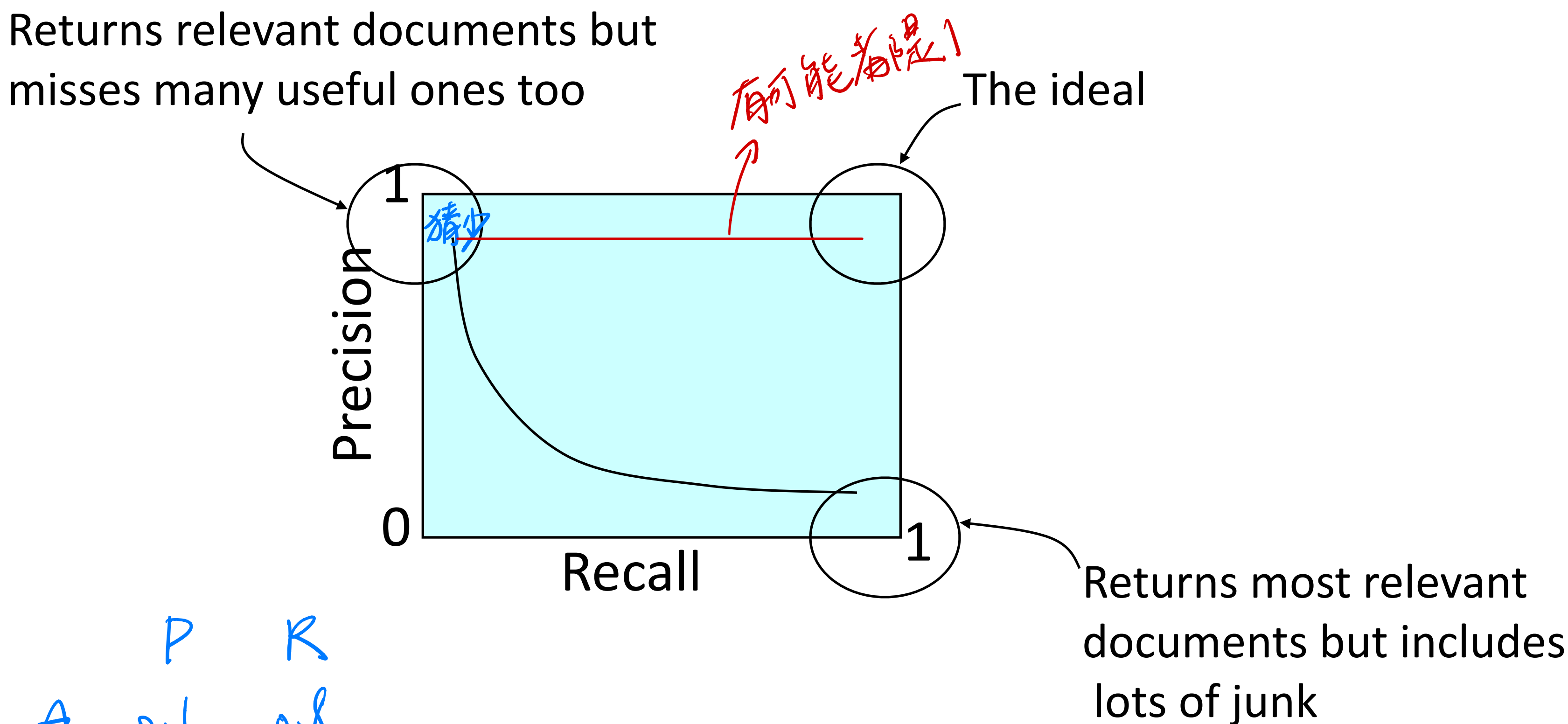
Why not just use accuracy?

for IR、不會用 accuracy

- How to build a 99.9999% accurate search engine on a low budget…. 沒有相關的文章 100 1000000 系統都不回答

- People doing information retrieval *want to find something* and have a certain tolerance for junk.

Trade-off between Recall and Precision



	P	R
IR A	0.6	0.8
IR B	0.7	0.5

$F_1 = \frac{2}{\frac{1}{P}+\frac{1}{R}}$ ⇝ 同時考慮 P、R，誰很少 ex 0.000…1 → ~0

F-Measure

	relevant	**not relevant**	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- $P = 20/(20 + 40) = 1/3$
- $R = 20/(20 + 60) = 1/4$
- $F_1 = 2 / (1/P + 1/R) = 2/7$

F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

E Measure (parameterized F Measure)

- A variant of F measure that allows weighting emphasis on precision over recall:

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R} + \frac{1}{P}}$$

- Value of β controls trade-off:
  - $\beta = 1$: Equally weight precision and recall (E=F).
  - $\beta > 1$: Weight recall more.
  - $\beta < 1$: Weight precision more.

EVALUATING RANKED RESULTS

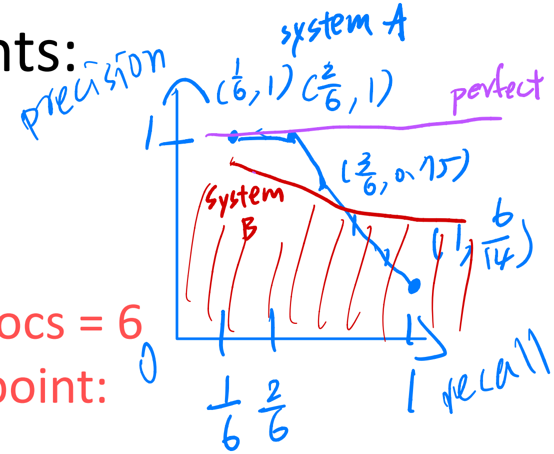
Evaluating ranked results

- Evaluation of ranked results:
  - The system can return any number of results
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*

Computing Recall/Precision Points

- For a given query, produce the ranked list of retrievals.
- Adjusting a threshold on this ranked list produces different sets of retrieved documents, and therefore different recall/precision measures.
- Mark each document in the ranked list that is relevant according to **the gold standard**.
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document.

Computing Recall/Precision Points: Example 1



n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	x

Let total # of relevant docs = 6

Check each new recall point:

R=1/6=0.167; P=1/1=1

precision recall curve

R=2/6=0.333; P=2/2=1

AUC

Area-Under-Curve

R=3/6=0.5; P=3/4=0.75

線底下的面積

R=4/6=0.667; P=4/6=0.667

誰面積大 → 愈好

R=5/6=0.833; p=5/13=0.38

R=6/6=1; p=6/14=0.41

Computing Recall/Precision Points: Example 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/3=0.667

R=3/6=0.5; P=3/5=0.6

R=4/6=0.667; P=4/8=0.5

R=5/6=0.833; P=5/9=0.556

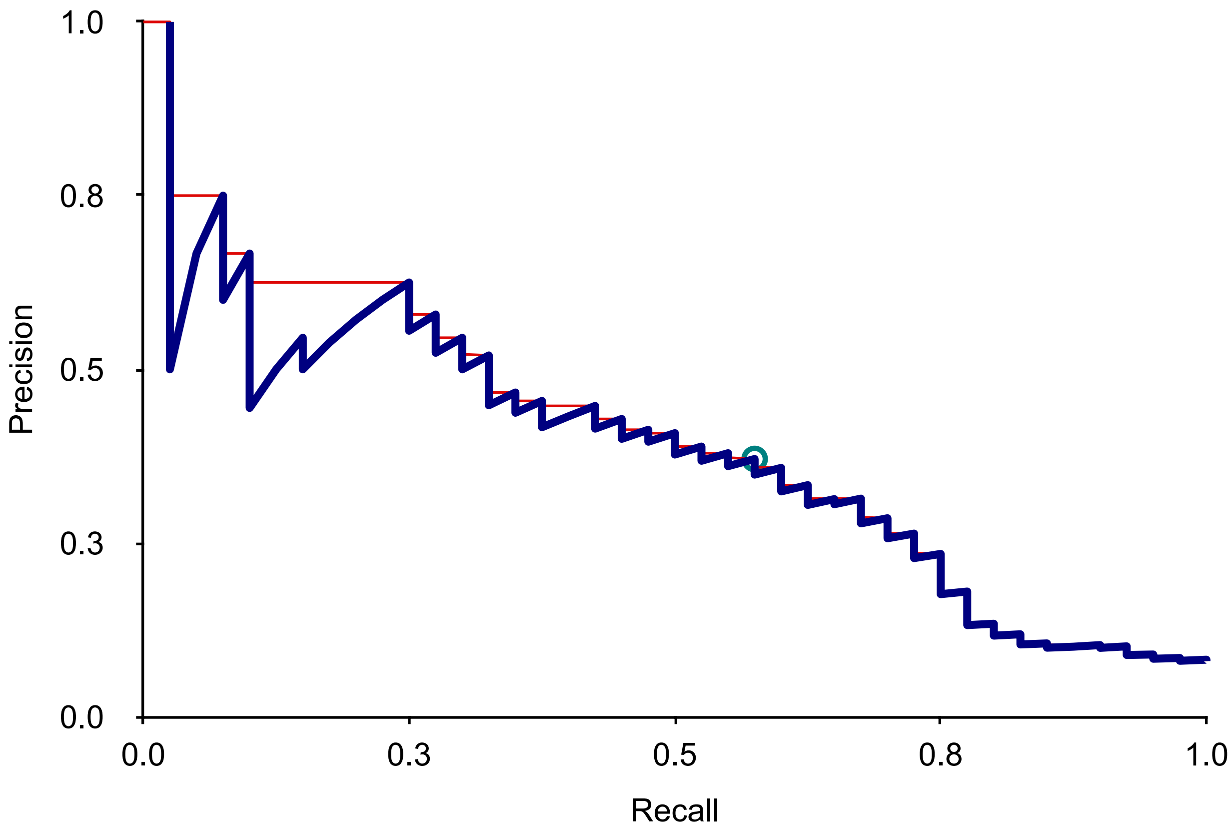
R=6/6=1.0; p=6/14=0.429

Interpolating a Recall/Precision Curve

- Interpolate a precision value for each *standard recall level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0$, $r_1 = 0.1$, …, $r_{10}=1.0$
- The interpolated precision at the *j*-th standard recall level is the maximum known precision at any recall level between the *j*-th and (*j* + 1)-th level:

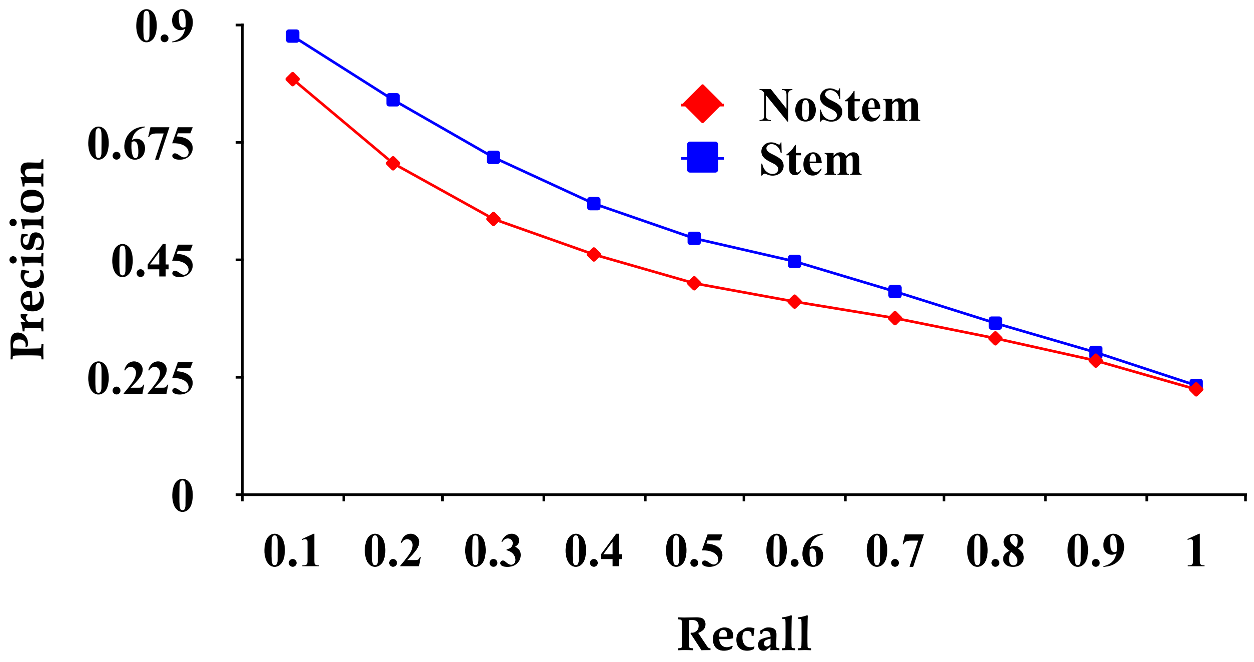
$$P(r_j) = \max_{r_j \leq r \leq r_{j+1}} P(r)$$

A precision-recall curve



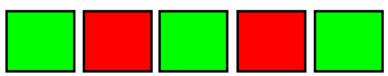
Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance



PRECISION@K

Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignores documents ranked lower than K
- Ex: 
  - Prec@3 of 2/3
  - Prec@4 of 2/4
  - Prec@5 of 3/5

How to choose k for precision@k measure ?

k怎麼設?

R- PRECISION

precision at R

relevant document 有幾個

R- Precision

- Precision at the R-th position in the ranking of results for a query that has R relevant documents.

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

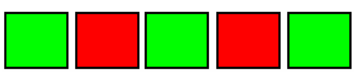
R = # of relevant docs = 6

R-Precision = 4/6 = 0.67

MEAN AVERAGE PRECISION

MAP

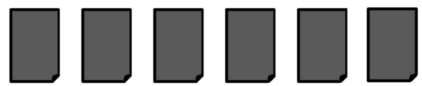
Mean Average Precision

- Consider rank position of each relevant doc
  - $K_1, K_2, \ldots K_R$
- Compute Precision@K for each $K_1, K_2, \ldots K_R$
- Average precision = average of P@K, for each K
- Ex:  has AvgPrec of $\frac{1}{3} \times \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{5} \right) \approx 0.76$

  $\frac{1}{3}\left(\frac{1}{1}+\frac{1}{3}+\frac{1}{5}\right)$?
- MAP is Average Precision across multiple queries/rankings

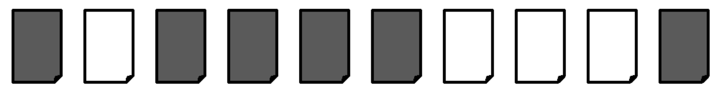
$$\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$$

Mean Average Precision



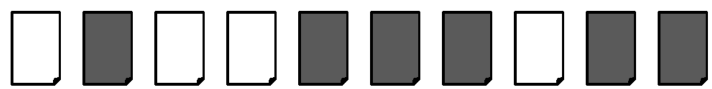
= the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

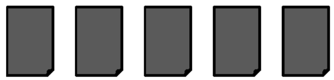


Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

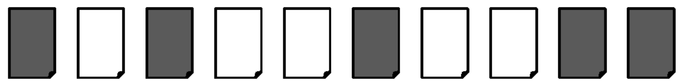
$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

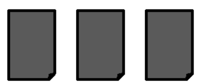
MAP


= relevant documents for query 1

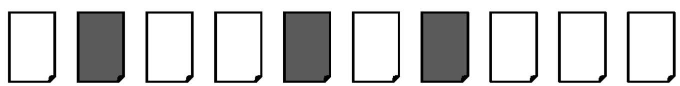
Ranking #1



Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5


= relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

$$average\ precision\ query\ 1 = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$

$$average\ precision\ query\ 2 = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$mean\ average\ precision = (0.62 + 0.44)/2 = 0.53$$

When there’s only 1 Relevant Document

- Scenarios:
  - known-item search
  - navigational queries
  - looking for a fact
- Search Length = Rank of the answer
  - measures a user’s effort

Mean Reciprocal Rank

RR

- Consider rank position, K, of first relevant doc

倒数

- Reciprocal Rank Score = $\frac{1}{K}$

第一名 → $\frac{1}{1}$ = 1

5 → $\frac{1}{5}$

- MRR is the mean RR across multiple queries

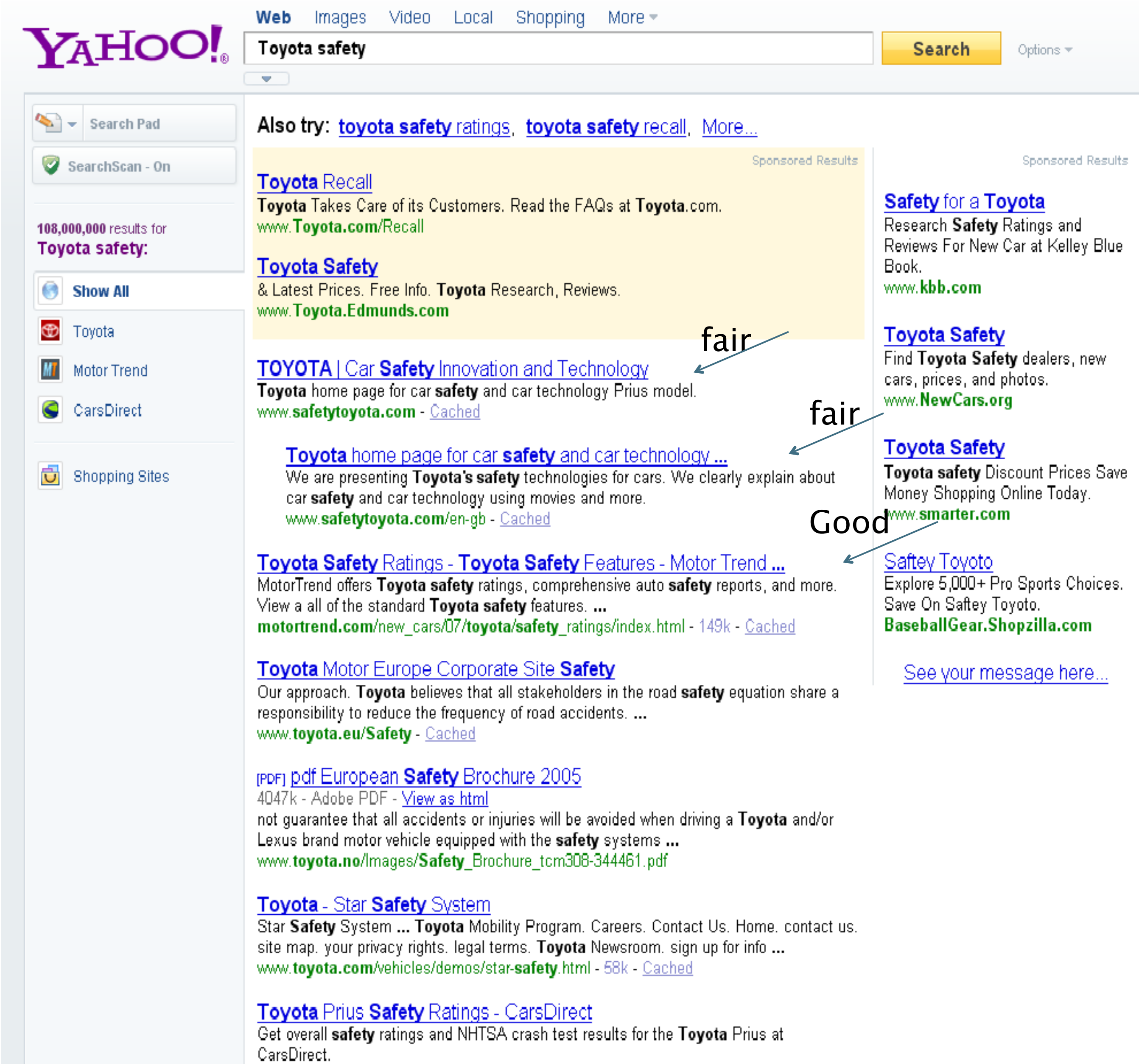
DCG

DISCOUNTED CUMULATIVE GAIN

相關程度致意

0、1、2、3……

權重值



YAHOO!

Web Images Video Local Shopping More

Toyota safety

Search

Options

Search Pad

SearchScan - On

108,000,000 results for **Toyota safety:**

Show All

Toyota

Motor Trend

CarsDirect

Shopping Sites

Also try: **toyota safety** ratings, **toyota safety** recall, More...

Sponsored Results

Toyota Recall
Toyota Takes Care of its Customers. Read the FAQs at **Toyota**.com.
www.**Toyota.com**/Recall

Toyota Safety
& Latest Prices. Free Info. **Toyota** Research, Reviews.
www.**Toyota.Edmunds.com**

TOYOTA | Car **Safety** Innovation and Technology
Toyota home page for car **safety** and car technology Prius model.
www.**safetytoyota.com** - Cached

Toyota home page for car **safety** and car technology ...
We are presenting **Toyota's safety** technologies for cars. We clearly explain about car **safety** and car technology using movies and more.
www.**safetytoyota.com**/en-gb - Cached

Toyota Safety Ratings - **Toyota Safety** Features - Motor Trend ...
MotorTrend offers **Toyota safety** ratings, comprehensive auto **safety** reports, and more. View a all of the standard **Toyota safety** features. ...
motortrend.com/new_cars/07/**toyota**/**safety**_ratings/index.html - 149k - Cached

Toyota Motor Europe Corporate Site **Safety**
Our approach. **Toyota** believes that all stakeholders in the road **safety** equation share a responsibility to reduce the frequency of road accidents. ...
www.**toyota.eu**/**Safety** - Cached

[PDF] pdf European **Safety** Brochure 2005
4047k - Adobe PDF - View as html
not guarantee that all accidents or injuries will be avoided when driving a **Toyota** and/or Lexus brand motor vehicle equipped with the **safety** systems ...
www.**toyota.no**/Images/**Safety**_Brochure_tcm308-344461.pdf

Toyota - Star **Safety** System
Star **Safety** System ... **Toyota** Mobility Program. Careers. Contact Us. Home. contact us. site map. your privacy rights. legal terms. **Toyota** Newsroom. sign up for info ...
www.**toyota.com**/vehicles/demos/star-**safety**.html - 58k - Cached

Toyota Prius **Safety** Ratings - CarsDirect
Get overall **safety** ratings and NHTSA crash test results for the **Toyota** Prius at CarsDirect.

Sponsored Results

Safety for a **Toyota**
Research **Safety** Ratings and Reviews For New Car at Kelley Blue Book.
www.**kbb.com**

Toyota Safety
Find **Toyota Safety** dealers, new cars, prices, and photos.
www.**NewCars.org**

Toyota Safety
Toyota safety Discount Prices Save Money Shopping Online Today.
www.**smarter.com**

Saftey Toyoto
Explore 5,000+ Pro Sports Choices. Save On Saftey Toyoto.
BaseballGear.Shopzilla.com

See your message here...

fair

fair

Good

Summarize a Ranking: DCG

- What if relevance judgments are in a scale of [0,r]? r>2
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)
  - $CG = r_1+r_2+\ldots r_n$

Rank	Score	
1	2	3
2	1	3
3	3	3
4	3	2
5	3	1

→ CG=12

can't 反應差別

CG = 2+1+3+3+3

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks

分: 愈高放愈前面

- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - the lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain

- Uses graded relevance as a measure of usefulness, or gain, from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks
- Typical discount is 1/log (rank)
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain (DCG)

- Discounted Cumulative Gain (DCG) at rank n
  - DCG = $r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots r_n/\log_2 n$
    - We may use any base for the logarithm, e.g., base=b

排名愈好 被扣分愈少

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:
  - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:
  - 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

    $\log_2 3$
  - = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:
  - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

DCG@k

希望0~1間 → NDCG

Discounted Cumulative Gain

- DCG is the total gain accumulated at a particular rank p:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - used by some web search companies
  - emphasis on retrieving highly relevant documents

Summarize a Ranking: NDCG

正規化

- **Normalized** Cumulative Gain (NDCG) at rank n
  - Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc
  - Compute the precision (at rank) where each (new) relevant document is retrieved => p(1),...,p(k), if we have k rel. docs
- NDCG is now quite popular in evaluating Web search

NDCG - Example

4 documents: d_1, d_2, d_3, d_4

i	Ground Truth		Ranking Function$_1$		Ranking Function$_2$	
	Document Order	r_i	Document Order	r_i	Document Order	r_i
1	d4	2	d3	2	d3	2
2	d3	2	d4	2	d2	1
3	d2	1	d2	1	d4	2
4	d1	0	d1	0	d1	0
	$NDCG_{GT}$=1.00		$NDCG_{RF1}$=1.00		$NDCG_{RF2}$=0.9203	

$$DCG_{GT} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4}\right) = 4.6309$$

$$DCG_{RF1} = 2 + \left(\frac{2}{\log_2 2} + \frac{1}{\log_2 3} + \frac{0}{\log_2 4}\right) = 4.6309$$

$$DCG_{RF2} = 2 + \left(\frac{1}{\log_2 2} + \frac{2}{\log_2 3} + \frac{0}{\log_2 4}\right) = 4.2619$$

$$MaxDCG = DCG_{GT} = 4.6309$$

Ad hoc results — Microsoft Research Ltd

Precion-Recall Curve

Summary Statistics	
Run Number	ok8amxc
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel-ret:	3212

$P = \frac{3212}{5000}$

Out of 4728 rel docs, we've got 3212

Recall=3212/4728

Recall Level Precision Averages	
Recall	Precision
0.00	0.8190
0.10	0.5975
0.20	0.5032
0.30	0.4372
0.40	0.3561
0.50	0.2936
0.60	0.2511
0.70	0.1941
0.80	0.1257
0.90	0.0696
1.00	0.0296
Average precision over all relevant docs	
non-interpolated	0.3169

recall precision curve

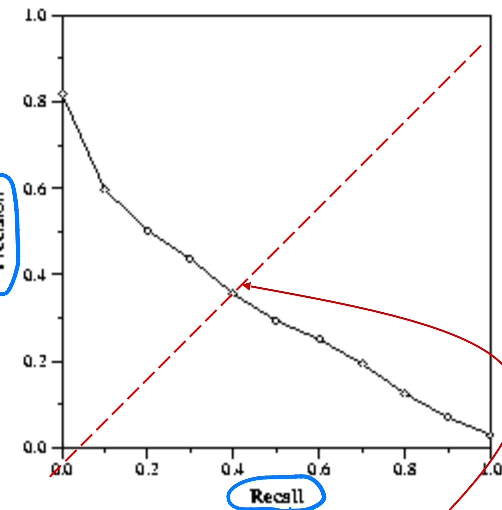
Mean Avg. Precision (MAP)

Document Level Averages	
	Precision
At 5 docs	0.5800
At 10 docs	0.5500
At 15 docs	0.4987
At 20 docs	0.4650
At 30 docs	0.4253
At 100 docs	0.2680
At 200 docs	0.1921
At 500 docs	0.1085
At 1000 docs	0.0642
R-Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3470

Precision@10docs

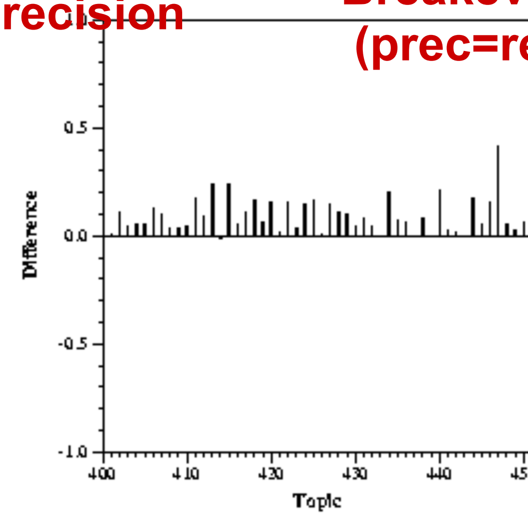
about 5.5 docs in the top 10 docs are relevant

R-Precision



Recall-Precision Curve

Breakeven Point (prec=recall)



Difference from Median in Average Precision per Topic

Recap

- Precision, Recall, Accuracy
- Recall-Precision Curve
- Precision@k
- R-Precision
- MAP Measure
- NDCG Measure